

Neil McGlohon, Ph.D.

CONTACT

neil@mcglo.dev
nmcglo.com

github.com/nmcglo
linkedin.com/in/nmcglo

SKILLS

Software Engineering: C/C++, Python, JavaScript/Node.js, HTTP API design, AMQP, SQL, Git, CMake, Jenkins, ArgoCD
Parallel & Distributed Computing: MPI, NCCL, SLURM, HPC systems, event-driven systems
AI/ML & Inference: PyTorch, TensorFlow, vLLM, Ray, AI agents, neuromorphic hardware
Cloud & Infrastructure: Kubernetes, OpenShift, Helm, Docker, Podman, Grafana, Prometheus, Infiniband & RoCE, AWS, S3, IBM Cloud, Longhorn, MetalLB, cloud-native systems, production-scale serving

PROFESSIONAL EXPERIENCE

– IBM, IBM Research

Almaden Research Center, San Jose, California USA

Staff Research Scientist

October 2022 – July 2026

- Led design and development of a purpose-built LLM inference engine for the IBM AIU NorthPole AI inference accelerator card, delivering state-of-the-art low latency and energy efficiency.
- Deployed rack-scale distributed AI systems of 288 NorthPole cards, generating over 500,000 tokens per second in aggregate while holding per-token latency as low as 1 millisecond.
- Led the team building the runtime stack and performance-critical SDK releases for NorthPole.
- Collaborated with model team to evaluate our quantized and retrained models using tools and benchmarks including EleutherAI's lm-eval and Hugging Face's OpenLLM leaderboards.
- Designed, built, and deployed a cloud-native OpenShift/Kubernetes stack that exposes and controls access to NorthPole-powered LLM inference appliances; integrated it into IBM watsonx, bringing NorthPole-accelerated LLM inference to workloads across the platform.
- Guided the team as subject-matter expert in parallel and distributed systems, containerization, orchestration, and LLM inference serving including vLLM plugin development and integration.
- Represented NorthPole at business development events, client outreach and showcases, presenting the platform to prospective partners. Bridged scientific and business-focused teams, driving knowledge transfer and cross-team collaboration.

– Rensselaer Polytechnic Institute, Center for Computational Innovations

Troy, New York USA

Research Scientist

July 2021 – September 2022

- Led software development and optimization of workloads on the Center's HPC systems for internal and partner research objectives.
- Served on the steering committee of an NSF-funded project-of-projects, mentoring participating graduate students as a subject-matter expert in HPC and AI systems.
- Collaborated with the IBM AI Hardware Center on research into distributed AI training benchmarking and simulation of composable systems.
- Triage and prioritized research-facilitation needs for thousands of users across multiple supercomputer systems.

Software Engineer

February 2020 – July 2021

- Maintained CODES, an open-source C/C++ framework for massively parallel simulation of high performance computing (HPC) communication networks, collaborating with researchers at Argonne and Sandia National Laboratories.
- Built custom software solutions in collaboration with external research partners.

– Cisco Meraki

San Francisco, California USA

Software Engineering Intern

May – August 2017

- Researched and developed firmware for Meraki's MR platform access points.
- Designed and built a lockless, thread-safe hashtable using RCU techniques, and upstreamed the contribution to Click, an open-source software-defined router project.

EDUCATION

- Ph.D., Computer Science** **July 2021**
 Rensselaer Polytechnic Institute, Troy, New York USA
 Research Focus: High Performance/Parallel Computing
- M.S., Computer Science** **May 2016**
 Rensselaer Polytechnic Institute, Troy, New York USA
 Research Focus: Distributed Computing, Machine Learning
- Bachelor of Science in Physics** **May 2014**
 University of Oklahoma, Norman, Oklahoma USA
 Minor: Computer Science

ACADEMIC
EXPERIENCE

- **Rensselaer Polytechnic Institute**, Troy, New York USA
Graduate Research Assistant **August 2014 – July 2021**
 M.S. and Ph.D. research, graduate-level coursework and projects. Areas of research: high performance and parallel computing, distributed systems, cloud computing, machine learning, neuromorphic computing and simulation.
Teaching Assistant: Computer Science I **August – December 2014**
 Supported the instructor of an introductory Python course: led lab sessions and discussions, held office hours, and proctored and graded exams/assignments.
- **University of Oklahoma**, Norman, Oklahoma USA
Undergraduate Research Assistant **May 2012 – May 2014**
 Explored electrical transport properties of quantum-confined antimony measured at cryogenic temperatures. Created and refined laboratory procedures for measuring differential conductance of a material, allowing for a deeper understanding of the interface between a topological insulator and a superconductor.
National Science Foundation REU **May – August 2012, 2013**
 Condensed matter research within NSF Research Experience for Undergraduates program including first observations of weak anti-localization in quantum-confined antimony films.

SCIENTIFIC
SERVICE

- NSF CAREERS Cyberteam Program
 Steering Committee Member (2020–2022)
- ACM SIGSIM Principles of Advanced Discrete Simulation (PADS)
 Organizing Committee, Ph.D. Colloquium Chair (2022)
 Organizing Committee, Publicity Chair (2023)
 Program Committee Member (2022–2025)
- ACM SIGPLAN Benchmarking in the Data Center: Expanding to the Cloud (BID)
 Program Committee Member (2022)
- IEEE IPDPS Scalable Deep Learning over Parallel and Distributed Infrastructure (ScaDL)
 Program Committee Member (2022)
- Peer-reviewer for 6 ACM/IEEE computing conferences and journals. (2019–Present)

PUBLICATIONS

- M. DeBole, R. Appuswamy, **N. McGlohon**, B. Taba, S. Esser, F. Akopyan, J. Arthur, A. Amir, A. Andreopoulos, P. Carlson, A. Cassidy, P. Datta, M. Flickner, R. Gandhasri, G. Garreau, M. Ito, J. Klamó, J. Kusnitz, N. McClatchey, J. McKinstry, T. Nayak, C. Tadeo Ortega Otero, H. Penner, W. Risk, J. Sawada, J. Sivagnaname, D. Smith, R. Sousa, I. Terrizzano, T. Ueda, T. Gray-Donald, D. Cox, and D. Modha. *A Scalable NorthPole System with End-to-End Vertical Integration for Low-Latency and Energy-Efficient LLM Inference*. arXiv 2511.15950v1 Pre-print article. November 2025
- R. Appuswamy, M. DeBole, B. Taba, S. Esser, A. Cassidy, A. Amir, A. Andreopoulos, D. Bablani, P. Datta, J. Kusnitz, N. McClatchey, **N. McGlohon**, J. McKinstry, T. Nayak, D. Smith, R. Sousa, I. Terrizzano, F. Akopyan, P. Carlson, R. Gandhasri, G. Garreau, N. Mimura Gonzalez, M. Ito, J. Klamó, Y. Nakamura, C. Tadeo Ortega Otero, W. Risk, J. Sawada, K. Schleupen, J. Sivagnaname, M. Stallone, T. Ueda, M. Flickner, J. Arthur, R. Panda, D. Cox, and D. Modha. *Breakthrough low-latency, high-energy-efficiency LLM inference performance using NorthPole*. IEEE High Performance Extreme Computing (HPEC), Boston, USA. September 2024.
- F. Akopyan, W. Risk, J. Arthur, A. Cassidy, M. DeBole, C. Tadeo Ortega Otero, J. Sawada, E. Colgan, M. Criscolo, P. Mann, H. Baier, K. Schleupen, A. Amir, A. Andreopoulos, R. Appuswamy, D. Bablani, P. Carlson, P. Datta, S. Esser, M. Flickner, R. Gandhasri, G. Garreau, M. Ito, J. Klamó, J. Kusnitz, N. McClatchey, **N. McGlohon**, J. McKinstry, Y. Nakamura, T. Nayak, J. Sivagnaname, D. Smith, R. Sousa, B. Taba, I. Terrizzano, T. Ueda, and D. Modha. *Breakthrough edge AI inference performance using NorthPole in 3U VPX form factor*. IEEE High Performance Extreme Computing (HPEC), Boston, USA. September 2024.
- E. Cruz, S. Qian, A. Shukla, **N. McGlohon**, S. Rakheja, C. D. Carothers. *Performance Evaluation of Spintronic-Based Spiking Neural Networks Using Parallel Discrete-Event Simulation*. ACM Transactions on Modeling and Computer Simulation (TOMACS), March 2024.
- E. Cruz, S. Qian, A. Shukla, **N. McGlohon**, S. Rakheja, C. D. Carothers. *Evaluating Performance of Spintronics-Based Spiking Neural Network Chips using Parallel Discrete Event Simulation*. ACM SIGSIM Conference on Principles of Advanced Discrete Simulation (PADS), Virtual. June 2022.
- N. McGlohon**, C. D. Carothers. *Toward Unbiased Deterministic Total Ordering of Parallel Simulations with Simultaneous Events*. ACM SIGSIM Winter Simulation Conference (WSC), Virtual. December 2021.
- N. McGlohon**, K. S. Hemmert, K. A. Brown, M. Levenhagen, S. Chunduri, R. B. Ross, C. D. Carothers. *Exploration of Congestion Control Techniques on Dragonfly-class HPC Networks Through Simulation*. IEEE SC'21 Workshop on Performance Modeling, Benchmarking and Simulation of High Performance Computing Systems (PMBS), Virtual. November 2021.
- J. Ma, J. Goodhue, K. Nelson, A. Sherman, E. Brown, C. D. Carothers, G. Collier, A. Del Maestro, A. Elledge, W. Figurelle, J. Huffman, G. Khanna, **N. McGlohon**, S. Najafi, J. Nucciarone, A. Schwartz, B. Segee, S. Valcourt, R. Zottola. *Leveraging Northeast Cyberteam Successes to Build the CAREERS Cyberteam Program: Initial Lessons Learned*. PEARC Workshop on Strategies for Enhancing HPC Education and Training (SEHET), Virtual. July 2021.
- K. A. Brown, **N. McGlohon**, S. Chunduri, R. B. Ross, E. Borch, C. D. Carothers, K. Harms. *A Tunable Implementation of Quality-of-Service Classes for HPC Networks*. ISC High Performance (ISC), Virtual. June 2021.
- N. McGlohon**, C. D. Carothers. *Unbiased Deterministic Total Ordering of Parallel Simulations with Simultaneous Events*. arXiv 2105.00069v1 Pre-print article. May 2021.
- N. McGlohon**, R. B. Ross, C. D. Carothers. *Evaluation of Link Failure Resilience in Multi-Rail Dragonfly-Class Networks*. ACM SIGSIM Conference on Principles of Advanced Discrete Simulation (PADS), Miami, USA. June 2020.
- N. McGlohon**, N. Wolfe, M. Mubarak, C. D. Carothers. *Fit Fly: A Case Study on Interconnect Innovation Through Parallel Simulation*. ACM SIGSIM Conference on Principles of Advanced Discrete Simulation (PADS), Chicago, USA. June 2019.

- Y. Kang, X. Wang, **N. McGlohon**, M. Mubarak, S. Chunduri, Z. Lan. *Modeling and Analysis of Application Interference on Dragonfly+*. ACM SIGSIM Conference on Principles of Advanced Discrete Simulation (PADS), Chicago, USA. June 2019.
- M. Mubarak, **N. McGlohon**, M. Musleh, E. Borch, R. B. Ross, R. Huggahalli, S. Chunduri, S. Parker, C. D. Carothers, K. Kumaran. *Evaluating Quality of Service Traffic Classes on the Megafly Network*. ISC High Performance (ISC), Frankfurt, Germany. June 2019.
- M. Plagge, C. D. Carothers, E. Gonsiorowski, **N. McGlohon**. *NeMo: A Massively Parallel Discrete-Event Simulation Model for Neuromorphic Architectures*. ACM Transactions on Modeling and Computer Simulation (TOMACS). September 2018.
- M. Plagge, **N. McGlohon**, C. Ross, C. D. Carothers. *Simulation and Visualization of Custom Neuromorphic Hardware using NeMo*. Neuromorphic Computing Symposium on Architectures, Models, and Applications, Oak Ridge National Laboratory, USA. July 2017.
- N. McGlohon**, S. Patterson. *Distributed Semi-Stochastic Optimization with Quantization Refinement*. American Control Conference (ACC), Boston, USA. July 2016.
- S. Patterson, **N. McGlohon**, K. Dyagilev. *Optimal k -Leader Selection for Coherence and Convergence Rate in One-Dimensional Networks*. IEEE Transactions on Control of Network Systems (TCNS). January 2016.
- S. Patterson, **N. McGlohon**, K. Dyagilev. *Efficient, Optimal k -Leader Selection for Coherent, One-Dimensional Formations*. European Control Conference (ECC), Linz, Austria. July 2015.
- S. Cairns, **N. McGlohon**, C. Robison, J. Keay, C. K. Gaspe, K. S. Wickramasinghe, T. D. Mishima, M. B. Santos, S. Q. Murphy *Weak anti-localization in ultrathin Sb(111) films*. American Physical Society, APS March Meeting, March 2013.